DOCUMENT RESUME

ED 186 482 TM 800 203

AUTHOR Humes, Ann
TITLE Assessing Writing Proficiencies with Writing Samples in Grades 2-6.
PUB DATE Apr 80
NOTE 17p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980). Parts may be marginally legible

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Education; Evaluation Criteria; *Scoring; Writing (Composition); *Writing Skills
IDENTIFIERS Analytic Scoring; Holistic Evaluation; Primary Trait Scoring; *SWPL Scoring System

ABSTRACT

Three different scoring systems are described; these systems are widely used for evaluating students' writing samples. A new model for scoring is presented by comparing and contrasting that model with the other systems. A pilot study of the new model for writing-sample items, scoring keys, and scoring guidelines is discussed. The pilot study was conducted in grades 2-6, and results suggest that the new model is appropriate for evaluating writing-samples that are a part of competency tests. Data from large-scale use on competency tests is also presented. (Author)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED186482

# Assessing Writing Proficiencies with Writing Samples in Grades 2 - 6

Ann Humes
SWRL Educational Research and Development
4665 Lampson Avenue
Los Alamitos, CA 90720

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Paper presented at
AERA
Boston, April, 1980

## ABSTRACT

Three different scoring systems are described; these systems are widely used for evaluating students' writing samples. A new model for scoring is presented by comparing and contrasting that model with the other systems. A pilot study of the new model for writing-sample items, scoring keys, and scoring guidelines is discussed. The pilot study was conducted in grades 2-6, and results suggest that the new model is appropriate for evaluating writing-samples that are a part of competency tests. Data from large-scale use on competency tests is also presented.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Ann Humes
Member, Professional Staff
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM800 203

Ann Humes
SWRL Educational Research
and Development
4665 Lampson Avenue
Los Alamitos, CA 90720

Paper presented at AERA,
Boston, April, 1980

# ASSESSING WRITING PROFICIENCIES WITH WRITING SAMPLES IN GRADES 2-6

This paper first describes three different scoring systems that are widely used for evaluating students' writing samples. Next it introduces a new model for scoring by comparing and contrasting the model with the three other systems. A pilot study of the new model for writing-sample items, scoring keys, and scoring guidelines used at grades 2-6 is then discussed. The study suggests that this scoring system is appropriate for evaluating writing samples that are included in competency tests at the elementary grades. Finally, data from large-scale use of the system are presented.

## SCORING SYSTEMS

The non-count scoring procedures frequently discussed and most widely used for evaluating writing samples are general impression (hereafter referred to as "holistic" scoring in accordance with this usage in much of the literature), analytic, and Primary Trait Scoring.

### Holistic Scoring

Holistic scoring involves a comparative evaluation of essays, as is performed by the Educational Testing Service (1976), the developers of the system. Readers are trained to read and rate papers in relation to other papers in a set. The papers are frequently scored on a scale ranging from one to eight (Mullis, 1975). Although a range of four may be used (ETS, 1976), some authorities believe that a range of eight or nine is optimal for making discriminations in the overall quality of writing when comparison is the basis for the scores (A Common Ground, 1978).

Holistic scoring is relatively simple. Some authorities laud the procedure as so simple that evaluators can merely sort a stack of essays into eight piles (Hogan and Mishler, 1979). Odell and Cooper (1978) hypothesize that the scoring ease is a significant reason for the method's wide acceptance. However, holistic scoring is frequently criticized because the only diagnostic information it offers is about the comparative quality of a paper--it does not indicate why a paper is a "4," or a "6," or an "8" (e.g., An Assessment, 1977; A Common Ground, 1978; Mullis, 1975).

### Analytic Scales

Analytic scoring is another frequently used technique. The definitive analytic scale was devised by Diederich (1966 and 1974). This scale consists of "General Merit" categories (the content characteristics) and "Mechanics" (form skills such as punctuation, capitalization, and grammar).

An analytic scale consists of a list of the prominent features of a particular writing type (e.g., expository discourse, narrative discourse). Thus it is designed to accommodate all tasks for that discourse type, regardless of differences in the stimuli eliciting the writing samples. Although five numerical ratings can be checked, only three ratings are described for each feature of the scale--low, middle, and high. When an essay is scored, the ratings on various features are tallied for subtotals of content and form scores and for a total writing score.

Many authorities advocate the analytic scale, asserting, for example, that it facilitates quick, efficient, and reliable scoring while providing diagnostic information (e.g., A Common Ground, 1978; Cooper, 1977). Research studies also support analytic scales (e.g., Smith, 1979; Pitts, 1979; Winters, 1979).

Analytic scales have critics as well as proponents. Cooper (with Odell, 1978) criticizes two underlying assumptions of the scale: (1) Criteria for rating good writing can be derived from sophisticated readers' perceptions of discourse types rather than from the specific writing task itself, and (2) the same criteria for judging a task elicited by one stimulus for a specific type of writing (e.g., expository) can be used to judge all tasks eliciting that writing type, even though the tasks are evoked by different stimuli.

## Primary Trait Scoring

Primary Trait Scoring (PTS) was devised by the National Assessment of Educational Progress (NAEP).

The primary trait of a writing task is determined by the purpose and audience for a specific piece of discourse and is rated on how well that "trait" is manifested. Each stimulus for a writing task is analyzed to identify the characteristics that effectuate that trait in the corresponding response--a writing sample; the characteristics (features) are unique to that stimulus, but not applicable to other stimuli for the writing type. These characteristics become the scoring criteria for the writing sample. Thus the scoring criteria for a writing task with an explanatory purpose will include some type of orderly presentation of specific information because "organization is usually the basis of the primary trait for a writing task requiring explanation" (Mullis, 1975, p. 11).

Although the analytic scale for expository (explanatory) writing also includes "Organization," as well as "Ideas" (factual content) as features to be scored, its accompanying scoring guidelines do not describe the specific organization or the type of factual content pertinent to the exact writing task. The Primary Trait Scoring system does.

Thus the important distinction of PTS is that specific writing tasks are evaluated--papers are not evaluated in relation to other papers in the set, as in holistic scoring, nor solely in relation to the general writing type, as in the Diederich analytic model. Furthermore, a paper can be excellent in terms of language, structure, and syntax, yet be rated as a failure if it does not exhibit the primary trait.

The stimulus for a writing task evaluated by PTS must be highly structured to facilitate scoring because the "more structured the task, the less difficult the scoring, since the essays will be more uniform in focus" (Mullis, 1975, p. 9). The corresponding scoring guide that is written for each stimulus describes the exercise in terms of characteristics displayed by a successful writing sample (Klaus, et al., 1979). From this scoring, examiners can develop a profile of students' writing abilities for the large elements of communication.

because PTS evaluation focuses on the primary trait of the writing task, the scoring is "independent of attention to mechanical and grammatical features" (Klaus et al., 1979, p. 23). When mechanics are concurrently scored, they are scored as the secondary or tertiary trait.

Definitive studies have not compared PTS with other scoring systems (Odell and Cooper, 1978). However, NAEP has reported on the field testing of its assessment, and PTS is being used by some states for their writing assessments (e.g., Kadomoto, 1979; Scherich, Liu, Kar, and Kettel, 1979).

Although the PTS system seems superior to many other non-count evaluation methods, its use for competency assessment can be criticized for several reasons. PTS does not evaluate grammar and mechanics except optionally as a secondary or tertiary trait; however, these features are commonly of interest to schools. Also questionable for competency testing is the PTS policy of assigning failure ratings to papers that are well written technically, but do not manifest the primary trait as described in the scoring guidelines. Lloyd-Jones (1977) cites a potential problem when scoring must be done by classroom teachers rather than professional readers, in his assertion that "In terms of convenience for the tester, Primary Trait Scoring is more difficult than other methods" (p. 40). Finally, PTS provides less specific diagnostic information than either analytic scales or the SWRL model discussed below.

## THE SWRL MODEL

The previous discussion suggests that the scoring system used to evaluate writing samples for competency tests should (1) be easy to score, thus requiring little or no training of evaluators, (2) be scored for the specific content features of the writing task rather than for the general writing type elicited, and (3) provide diagnostic information. Additionally, form skills should be assessed because they

are important to writing competency (Rentel, King, and Pappas, 1979); furthermore, performance on form skills may influence evaluators' scoring of content (either negatively or positively) if the scoring system does not provide for their separate evaluation.

The SWRL scoring system addresses these considerations. It combines the simplicity of the analytic scales, thus being easy to score, with the specificity of Primary Trait Scoring. Each highly structured stimulus also facilitates easy scoring because it is always accompanied by scoring criteria that pertain specifically to that stimulus. For an example of a writing sample stimulus, see Figure 1. For the corresponding scoring key, see Figure 2.

As illustrated by the scoring key in Figure 2, the skills array of the scoring matrix consists of the features of good writing that the sample exhibits when a student employs the content and form skills appropriate for both the specific writing task and the grade level of assessment. Thus an explanatory writing sample that gives directions, as the example in Figure 1, should sequentially present specific instructions. Also, this information should be conveyed in precise language. The SWRL scoring key displayed in Figure 2 typically simplifies the evaluation of these characteristics by describing how they are specifically implemented for this individual task:

| Trait | Implementation |
|---|---|
| Specific Information: | Includes all parts of the treat. |
| | Uses sentences that pertain to the main idea (this criterion also assesses coherence). |
| Sequential presentation: | Writes about the steps in the order in which they are performed. |
| | Uses logical transitions (this criterion also assesses coherence). |
| Precise language: | Uses precise language. |

This specificity contrasts sharply with the general criteria for evaluating an expository (explanatory) task in the analytic model.

In contrast with the holistic system, SWRL scoring provides considerable diagnostic information. In fact, SWRL scoring provides more detailed diagnostic information than does PTS. For PTS, various aspects of content are collapsed into one category, the primary trait, and thus are given a single point score for the degree to which this overall trait is exhibited. Thus PTS does not provide diagnostic information for individual components (e.g., evocative details, transitions, topic sentences) of the primary trait. When form skills

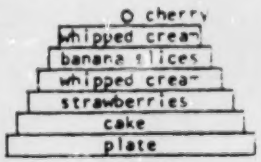
Directions: Read the contents and study the layers of the Fruit and Whipped Cream Treat.

## Fruit and Whipped Cream Treat

**Contents**

1 yellow cake
1 cup of strawberries
1 cup of cut-up bananas
2 cups of whipped cream
1 cherry

Specifications for picture to be drawn:




Write a paragraph that explains to someone who has never fixed food how to put the Fruit and Whipped Cream Treat together.

- Start with the first thing you do.
- Tell about each part of the Treat that you put together.
- Tell about the parts in the order that you must put them together.
- Be sure to use exact words.
- Use words like next to join some of the sentences in your paragraph.

Figure 1. SWRL Writing Sample Stimulus.

| Scoring Criteria | Good | Acceptable | Unacceptable |
|---|---|---|---|
| CONTENT: | | | |
| Includes all parts of the Treat. | ___ | ___ | ___ |
| Writes about the steps in the order in which they are performed. | ___ | ___ | ___ |
| Uses precise language. | ___ | ___ | ___ |
| Uses logical transitions. | ___ | ___ | ___ |
| Uses sentences that pertain to the main idea. | ___ | ___ | ___ |
| FORM: | | | |
| Uses correct grammar. | ___ | ___ | ___ |
| Uses good sentence structure. | ___ | ___ | ___ |
| Capitalizes correctly. | ___ | ___ | ___ |
| Punctuates correctly. | ___ | ___ | ___ |
| Spells correctly. | ___ | ___ | ___ |
| Writes legibly, with appropriate margins and indentation. | ___ | ___ | ___ |

Figure 2. SWRL Scoring Key for Figure 11.

BEST COPY AVAILABLE

such as grammar and mechanics are assessed as the secondary or tertiary trait in PTS, these skills are also collapsed for a single score so that, for example, a score of two on the tertiary trait would not indicate whether errors reflected problems in punctuation, capitalization, or other conventions, or of a combination of these categories.

In contrast with PTS, the SWRL model lists content and form features for separate scoring, and performance ratings of good, acceptable, or unacceptable are given on each criterion. Thus the scoring elicits considerable diagnostic information about the individually listed component skills. Also in contrast with PTS (and consonant with analytic scales), the system provides a total score for the writing sample when numerical equivalents are assigned to the good/acceptable/unacceptable ratings of the content and form skills and these individual scores are tallied. Furthermore, the SWRL model does not prescribe the numerical values to be assigned to the good/acceptable/unacceptable categories as does the analytic scale; this "omission" allows users of the scoring system to assign weighted values that correspond to the importance of the skills to the population to be assessed.

As with the other systems, scoring guidelines (see Figure 3) accompany the stimulus; these guidelines describe the features that constitute a good, acceptable, or unacceptable score on each criterion in the scoring key. In contrast with PTS, the guidelines are simpler to interpret and thus to use because they correspond to precise individual features that implement the whole "primary trait." In contrast with analytic scales (and consonant with PTS), the guidelines are specific to the task described in the writing-sample stimulus rather than to any stimulus for a particular writing type.

Scoring is further simplified in the SWRL model by a range that spans three broad rating categories. Although authorities differ over the number of ratings a scoring range should span, they generally agree that the fewer the possible rankings, the simpler the scoring. Holistic scales may frequently entail as many as eight or nine points (A Common Ground, 1978) or as few as four (ETS, 1976), but these numbers are for an entire essay rather than for individual criteria.

At the other extreme from the holistic scoring range is the dichotomous scale. This scale consists of only two ratings for each feature/skill to be evaluated--acceptable/unacceptable or yes/no. Cooper (1977) doubts that dichotomous scales yield reliable scores on individuals. Additionally, dichotmous scales tend to place a disproportionate number of students into the unacceptable range, and this phenomenon is inappropriate for competency testing. In informal discussions of rating scales, evaluators have reported that decision-making is difficult in dichotomous scoring since superior papers must be assigned the same acceptable/yes ratings as mediocre papers. Because of the wide disparity in performance, scorers tend to rate borderline papers as unacceptable in comparison with superior papers.

Includes all parts of the Treat (i.e., plate, cake, strawberries, whipped cream, banana slices, whipped cream, cherry).

Good: All parts are included.

Acceptable: Most parts are included.

Unacceptable: Many parts are missing.

Writes about the steps in the order in which they are performed (i.e., cake on plate, strawberries on cake, whipped cream on strawberries, banana slices on whipped cream, whipped cream on banana slices, cherry on whipped cream/cherry on top).

Good: Most or all steps are in the right order.

Acceptable: Many steps are in the right order.

Unacceptable: Many steps are not in the right order.

Comment: If fewer than three steps are included, do not score this criterion.

Uses precise language.

Good: Precise words are generally used to explain the procedure; e.g., an exact word like "spread" is used to explain adding a food part to the top of a previous food part.

Acceptable: General words are used to explain the procedure; e.g., "put on" is used to explain adding a food part to the top of a previous food part.

Unacceptable: Language used does not convey a sense of the procedure.

Uses logical transitions.

Good: Several logical transitions for time (e.g., "then," "next"), space (e.g., "on top of"), and possibly reason (e.g., "because") are used.

Acceptable: Few logical transitions are used.

Unacceptable: No logical transitions are used.

Figure 3. Sample of NAEP Scoring Guidelines

A four-point scale is used by NAEP in Primary Trait Scoring: "The first and lowest score point is for respondents who show no evidence of the skill being measured. The second point is defined as some evidence of the skill. Score point '3' is defined as demonstrating competence, and score point '4' is reserved for those papers which are excellent or outstanding" (Mullis, 1975, p. 10). Educational Testing Service (ETS) also uses a four-point scale because "readers are forced away from a middle, or uncommitted, score" (ETS, 1976, p. 3).

Forcing scores away from the middle range, as is done by both NAEP and ETS, is inappropriate for minimal competency testing. A four-point scale implies that two is below average, and inevitably, more writing samples will be perceived as below average because the scale has two categories below the midpoint of the range.

Five-point scales are sometimes used for analytic scoring in accordance with Diederich's model. Although the scales span five points, the corresponding guidelines describe only three ratings: low, middle, and high (Diederich, 1974). No guidelines are provided to aid scorers in discriminating between 1 and 2 or between 4 and 5. Further, Lloyd-Jones (1977) criticizes the five-point scale, asserting that "'3' is not the midpoint of a five-point scale; in this rating '5' is exceptionally good. '1' and '2' represent failures. . . '3' and '4' represent degrees of success. '5' is a rarity of success. . ." (p. 50).

For the reasons above, only a three-point scale was acceptable for the SWRL model. Because there is only one category below the midpoint, fewer papers are likely to be scored as below average with this scale than with scales of two, four, or five. Additionally, the broad discriminations evaluators must make do not entail so much information processing as do systems with wider ranges: Wider ranges require evaluators to remember too many distinctions for the different ratings of all the features to be scored.

## THE PILOT STUDY

Writing-sample items and scoring keys were to be included on the assessment instruments for the grade-by-grade competency testing in a large urban school district. Prior to the administration of the assessments in the spring of 1979, a pilot study was conducted on writing sample items parallel to those to be included at grades 2-6. The pilot study was designed to determine (1) the appropriateness of the composing tasks, (2) the utility of the scoring keys, particularly in relation to their effectuating consistency and thus objectivity in the scoring, (3) time requirements for scoring writing samples when a scoring key is used, and (4) the efficacy of scoring guides as aids for evaluating writing samples.

### Materials

A writing-sample stimulus was prepared for each grade level. These stimuli were designed to parallel actual items developed for the assessment instruments. Each writing stimulus was designed to elicit a specific discourse type appropriate for students in the corresponding grade level. Grade levels of the writing tasks were determined on the basis of data from detailed analysis of language-arts textbooks for elementary schools (Humes, 1978). The second-grade task elicited narrative discourse; students were to write sentences for pictures that together, in sequence, suggest a story. Descriptive discourse was evoked for the third-grade task, which asked students to describe a sensory object. The fourth-, fifth-, and sixth-grade items were designed to elicit expository discourse: fourth grade, a newspaper account; fifth grade, map directions; sixth-grade, a contrast of two elements of the same classification. A scoring key and a scoring guide were developed for each writing-sample stimulus.

## Procedures

The pilot study was conducted in February, 1979. Materials were mailed to the school for teachers to administer at their convenience to their classes. Teachers were directed to allow students ample time to complete the writing samples, but the time was not to exceed 45 minutes.

Six samples from each grade level were used for scoring; they were selected randomly with the exception of two excluding criteria: (1) Samples composed in a language other than English were not selected; (2) although the scoring keys list legibility as an evaluated criterion, illegible samples were not selected because their unreadability precludes evaluating student performance on the other criteria.

The samples were scored on corresponding scoring keys by four evaluators (EG1): SWRL staff experienced with elementary school writing.

When all evaluators had scored all samples for a grade level, numerical values were applied to the scoring categories. Then scores were tallied, giving each writing sample four total scores, one from each evaluator. The data were then analyzed across samples and across evaluators.

On the bases of the analysis and evaluators' comments, some of the scoring keys were slightly revised. Then four of the five sets* of the writing samples were distributed for scoring by the second group of four evaluators (EG2): SWRL staff with little or no experience with elementary school writing. All procedures but one were the same as for EG-1 above. The one varying procedure was that scoring guidelines for each set of writing samples were given to each member of EG-2 to help determine good/acceptable/unacceptable ratings.

## Results

Appropriateness of the Tasks. The numerical equivalents assigned to the scores facilitated determining the appropriateness of the composing tasks on the basis of the students' success. Four of the five composing tasks were appropriate because nearly all samples received acceptable or better total scores. On those four sets, evaluators in EG-1 gave only 5% "unacceptable scores." Evaluators in EG-2, who received the scoring guides to facilitate evaluation, gave only 1% unacceptable scores.

---

*Reasons for omitting one writing sample are given in the results section.

The fourth grade writing sample item proved to be inappropriate. It directed students to write a newspaper account from listed "reporter's notes." Students received more unacceptable and marginal scores on this item than on any other composing task. The kinds of problems evidenced in the samples suggest that students may not be familiar with the important features of writing newspaper accounts despite the high frequency of this task in students' textbooks at both third and fourth grades.

Because of the problems posed by this pilot-study item, the parallel item in the actual assessment instrument was replaced by an item for writing a personal letter. Consequently, evaluators in EG-2 were not asked to score this set of writing samples.

Utility of the Scoring Keys. In general, the scoring keys facilitated rapid scoring of the writing samples (see time section below) on a consistent and objective basis. Although some evaluators gave scores that were higher or lower than scores assigned by other evaluators, they tended to be fairly consistent within themselves. Thus their rank order for the samples generally matched that of the other scorers. In cases for which rank order varied, evaluator error was usually responsible.

In cases of scorer error for EG-1, the error generally resulted from misinterpretation of the scoring key; this problem can be precluded by providing well-written scoring guidelines. For example, ratings on the spelling criterion varied for EG-1; some evaluators scored students low for misspelling the same word many times, while other evaluators scored the repeated error as one misspelling. The scoring guidelines ameliorated this problem.

Corresponding with total score consistency, evaluators also tended to agree on the scoring of specific criteria on individual papers. (Agreement was determined by looking at the scores given by each evaluator for each sample for each criterion on the scoring key.) For 47% of the criterion ratings all evaluators agreed completely; for 48%, they disagreed by only one degree (i.e., good/acceptable or acceptable/unacceptable); for 5%, they spread their scorers across the three degrees. Again, this discrepancy resulted, in most such cases, from evaluator error.

Time Required to Score Writing Samples. Scoring the writing samples was not a time-consuming process. The mean time required to score writing samples at grade 2 was only 1.95 minutes, with a range of 1.17 to 2.67 across scorers; the mean time for the process with grade 6 papers was 2.8 minutes, with a range of 1.67 to 3.33. Mean times for other grades were between these ranges. Further, an accelerated rate of scoring can be anticipated in large-scale assessment because an evaluator will become increasingly familiar with the single task of the grade-level assessment and with the evaluation criteria of the one scoring key.

The Efficacy of the Scoring Guidelines. To determine whether the benefits of scoring guidelines may be negated by the time required to read them, the evaluators who used the guidelines (EG2) recorded preparation time. The following are the average times required for familiarization with the scoring guide:

Grade 2: Three minutes
Grade 3: Two minutes
Grade 5: Nine minutes
Grade 6: Five minutes

These averages suggest that the time required to study the scoring guide is not excessive if it facilitates more equitable and objective scoring. This equity and objectivity is suggested by the fact that the range of scores across evaluators for each writing-sample was generally lower for EG-2 than for EG-1 scores.

## USE ON COMPETENCY ASSESSMENTS

Competency tests that included the SWRL writing samples were administered in the spring of 1979. Writing-sample items were included for grades 2-6, but samples were required only for the third and sixth grade classes. Approximately 76,000 students completed the items.

Sums of the "good" and "acceptable" categories give overall scores on the writing-sample items that are higher than scores on the multiple-choice items, with the exception of those for second grade. Totals of all three categories (good/acceptable/unacceptable) add up to more than 100% for all grades but fifth because of the uneven number of scoring categories: Some criteria had only two possible ratings; e.g., originality could only be scored "good" or "acceptable." The following table compares the overall scores on the multiple-choice items with the scores on the writing samples.

| | | Writing Sample | | | |
|---|---|---|---|---|---|
| | Multiple-Choice | Percent Good | Percent Acceptable | Percent Unacceptable | "Good" and "Acceptable" Summed for Overall Score |
| Grade 2 | 77% | 47 | 28 | 30 | 75 |
| Grade 3 | 79% | 55 | 30 | 21 | 85 |
| Grade 4 | 69% | 36 | 37 | 30 | 73 |
| Grade 5 | 63% | 28 | 41 | 31 | 69 |
| Grade 6 | 57% | 29 | 44 | 29 | 73 |

Only a small percentage of students completed the writing samples at grades two, four, and five, so the scores at these grades might have been different if more students had completed the tasks. Problems with specifics of the writing-sample stimulus depressed scores at both fifth and sixth grades. For example, low scores were obtained on the capitalization criterion at fifth grade: however, the map that served as the stimulus included capitalized labels for landmarks (e.g., "Supermarket" "Stoplight"). The sixth grade stimulus was designed to elicit a contrast of abstractions rather than of pictured and concretely describable entities in order to assess exposition rather than sensory description; however, a picturable stimulus would have been more motivating and contrastable.

## CONCLUSION

The careful structuring of a complete assessment package (i.e., stimulus, scoring key, and scoring guidelines) is an exacting and time-consuming task. However, maximal effort put into the development of the package results in minimal effort for the users. Evaluating writing samples is a comparatively simple task that can be quickly expedited by the classroom teacher when the SWRL system is used. The model presented here has proven to be appropriate and efficient for eliciting and evaluating writing samples administered to determine the writing competency of elementary school students. Furthermore, the diagnostic information provided by this system can be used to ameliorate instructional inadequacies and assure student learning. Such results will go a long way toward allaying professional concern about competency testing.

## REFERENCES

A common ground for assessing competence in written expression. Publication of the Office of the Los Angeles County Superintendent of Schools, Division of Curriculum and Instructional Services, 1978.

An assessment of the writing performance of California high school seniors. Sacramento: California State Department of Education, 1977.

Cooper, C. R. Holistic evaluation of writing. In Cooper, C. R., and Odell, L., eds., Evaluating Writing: Describing, measuring, judging. Buffalo: State University of New York, 1977, pp. 3-31.

Diederich, P. B. How to measure growth in writing ability. English Journal, 1966, 55, pp. 435-449.

Diederich, P. B. Measuring growth in English. Urbana, IL: National Council of Teachers of English, 1974.

Educational Testing Service. How the essay in the CEEB English Composition test is scored: An introduction to the reading for readers. Princeton: Educational Testing Service, 1976.

Hogan, T. P. & Mishler, C. J. Judging the quality of student's writing: When and how. The Elementary School Journal, 1979, 79:3, pp. 142-146.

Humes, A. Proficiency Verification Systems (PVS): Skills indices for language arts (Technical Note No. 3-78-14). Los Alamitos, California: SWRL Educational Research and Development, 1978.

Kadomoto, L. Hawaii writing assessment. Unpublished 1979 document received by personal communication with L. Kadomoto, Evaluation Specialist for the Department of Education for the State of Hawaii.

Klaus, C. H., Lloyd-Jones, R., Brown, R., Littlefair, W., Mullis, I., Miller, D., & Verity, D. Composing childhood experience: An approach to writing and learning in the elementary grades. St. Louis: CEMREL, Inc., 1979.

Lloyd-Jones, R. Primary trait scoring. In Cooper, C. R. & Odell, L., eds., Evaluating writing: Describing, measuring, judging. Buffalo: State University of New York, 1977, pp. 33-66.

Mullis, I. V. S. The primary trait system for scoring writing tasks. Denver: National Assessment of Educational Progress, 1975.

National Assessment of Educational Progress. Report 3: 1969-1970 writing: National results. Denver: National Assessment of Educational Progress, 1970.

Odell, L. & Cooper, C. Evaluating procedures for evaluating writing. Paper presented at AERA, Toronto, March, 1978.

Pitts, M. The measurement of students' writing performance in relation to instructional history. Paper presented at AERA, San Francisco, April, 1979.

Rentel, V. M., King, M., and Pappas, C. C. Toward a theory of early writing development. Paper presented at AERA, San Francisco, April, 1979.

Scherich, H. H., Liu, S., Kar, R., & Kettel, R. The development of writing production assessment items and scoring systems for the Louisiana State Pupil Assessment Program. Paper presented at AERA, San Francisco, April, 1979.

Smith, L. S. Measures of high school students' expository writing: Direct and indirect strategies. Paper presented at AERA, San Francisco, April, 1979.

Winters, L. Alternative scoring systems for predicting criterion group membership. Paper presented at AERA, San Francisco, April, 1979.